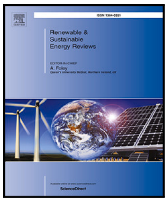
Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

journal homepage: www.elsevier.com/locate/rser



Sub-minute probabilistic solar forecasting for real-time stochastic simulations

Dazhi Yang [a,*], Gokhan Mert Yagli [b], Dipti Srinivasan [c]

[a] *School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin, Heilongjiang, China*
[b] *Solar Energy Research Institute of Singapore, National University of Singapore, Singapore*
[c] *Department of Electrical and Computer Engineering, National University of Singapore, Singapore*

ARTICLE INFO

Keywords:
Solar forecasting
Stochastic simulations
Lasso
Analog ensemble
Probabilistic forecasting

ABSTRACT

Simulation needs to reflect reality, otherwise, it yields misleading results that are potentially harmful to the operators and decision makers who rely on that simulation. Current stochastic simulation methods for solar energy systems in smart grids mostly consider scenarios generated from a single low-frequency time series, which is unable to reflect the high-frequency fluctuation and changing uncertainty in the actual solar power output. To that end, this paper introduces a state-of-the-art probabilistic solar forecasting method, that provides remedies to the drawbacks, and thus benefits real-time stochastic simulations at large. Lasso-penalized quantile regression is paired with an analog-based preselection algorithm, and is used to forecast the irradiance over a 1 km by 1 km area, in sub-minute timescales. To capture the changing weather condition, the forecasting method uses online training, and updates its parameters and forecasts every few seconds. Despite the heavy computation required, the forecasting method is fast; it takes less than 1 s to complete a forecasting cycle. Through five benchmarking methods, ranging from the naïve climatology and persistence to the state-of-the-art analog ensemble (AnEn), the proposed method is shown to be able to attain an exceptionally high forecast skill in terms of pinball loss (up to a skill score of 55%, with AnEn being the reference model), which is unparalleled by all previous works that used the same dataset. To promote the future uptake of the method, the R code and the final forecast datasets are released on Github.

1. Introduction

Coordinating generation from multiple energy sources, storage, and load has become an integral part of energy management systems ubiquitously used in smart grids. Because the actual load and renewable generation can deviate from forecasts, adequate uncertainty quantification is necessary in order to ensure reliable and stable operations [1]. Whereas stochastic simulation is no doubt an appropriate tool to study the uncertainty propagation of renewable generation and its grid implications, the precise simulation requirements, for example, whether or not additional balancing efforts are needed, become increasingly opaque when the simulation time scale approaches real time, that is, of the order of seconds, or sub-minute [2,3].

That said, despite the importance of sub-minute stochastic simulation, most existing works focus on hourly or day-ahead scenarios [e.g., 4–7]. On this point, an excellent justification for the necessity of sub-minute renewable generation forecasting can be found in [3]. In that, the authors discussed the problem of secondary control in power systems, particularly the automatic generation control (AGC), in which the reserves may not follow the rapid and large ramps introduced by renewables tightly enough. However, in the simulation part of the work, only some naïve hypothetical point (i.e., deterministic) forecasts are used, which might be too ideal to represent the actual forecast uncertainty involved. Aside from secondary control, sub-minute forecasts are also needed for decision-making in photovoltaic (PV) plant control, e.g., curtailment strategy or energy storage control, as detailed in [8,9]. In that work, solar irradiance was not measured, but was inferred from PV-array measurements, which might influence the subsequent forecasts negatively. Indeed, state-of-the-art solar forecasts are rarely used in stochastic simulation works, even when the time scale is hourly or daily [10].

Abbreviations: AGC, automatic generation control; AnEn, analog ensemble; CRPS, continuous ranked probability score; ECDF, empirical cumulative distribution function; lpQR, lasso-penalized quantile regression; MLR, multiple linear regression; NREL, National Renewable Energy Laboratory; OSMG, Oahu Solar Measurement Grid; PIAW, prediction interval average width; PICP, prediction interval coverage probability; PeEn, persistence ensemble; PV, photovoltaic; VAR1, vector autoregressive model of order 1

[*] Corresponding author.
E-mail address: yangdazhi.nus@gmail.com (D. Yang).

https://doi.org/10.1016/j.rser.2021.111736
Received 30 November 2020; Received in revised form 12 September 2021; Accepted 30 September 2021
Available online 20 October 2021

The possible reason for the under-utilization of sub-minute solar forecasts is two-fold. Firstly, there seems to be a disconnect between the solar energy community and the power system community. As the state-of-the-art solar forecasting advances at an unprecedented speed [11], many algorithms and methods are being proposed on a weekly basis, and it is generally difficult to identify the best methods, since all papers seem to claim superiority over others [12]. Secondly, and more importantly, in spite of the importance of solar forecasting, it is in fact not the core focus of the power system community, as compared to other topics such as load flow, frequency control, or demand-side management. Hence, many researchers naturally put more emphasis on investigating the core research topics at hand, and forecasting is but one pre-requisite.

To address the situation, an obvious solution is to introduce open-source datasets and methods that are ready-to-use. However, claiming a method "good enough" or "sufficiently accurate" is not a straightforward task. In that, one ought to consider the nature of the solar forecasting problem. Stated differently, before any standardized forecast dataset is prepared, one needs to know the desirable characteristics and salient features of the state-of-the-art solar forecasting. Nonetheless, to fully describe and argue all desirable characteristics and salient features, a full-length review is required. Hence, this paper only gives brief justification for each characteristic, and the readers are referred to the references for more details.

Firstly, good solar power forecasts are rarely generated using power measurements directly. Instead, one ought to first forecast the irradiance and then convert that to solar power in the second step [11,13,14]. This is because irradiance is spatio-temporal in nature, forecasts generated using information from a single location (i.e., the PV farm location) are unable to capture the cloud dynamics, which is the main cause of variability in solar power output; spatio-temporal forecasts are therefore preferred [15,16]. Secondly, a weather forecast is intrinsically five-dimensional, spanning space, time, and probability [17,18]. Probabilistic forecasts offer rigorous uncertainty quantification, and carry more informative than deterministic forecasts. Although the current operational power system control and scheduling rarely take probabilistic solar forecasts into consideration, numerous works have advocated the necessity of moving from deterministic to probabilistic (or stochastic) operations [e.g., 19,20]. Thirdly, the forecasting algorithm needs to be computationally efficient and fast, which is particularly true for sub-minute forecasting, where model parameters are refitted or re-optimized every few seconds. Here, the well-known benefits of online forecasting justify the need for forecasting with moving windows [21,22]. Lastly, forecast quality depends on how data are handled. Solar irradiance and PV power time series is known to be nonstationary, and clear-sky model is the best known method to reduce such nonstationarity, and hence is almost always used during solar forecasting [23]. On top of that, stringent but not excessive quality control is also vital to the robustness of the forecasts generated [24,25].

In view of the above discussion, to bridge the gap between sub-minute grid simulations and solar forecasting, considerable scientific effort and resources should be invested in developing novel approaches. This paper proposes a sub-minute probabilistic solar irradiance forecasting method with online training. Data from a dense sensor network that measures high-resolution irradiance are considered. Sensor networks allow a forecaster to utilize spatio-temporal information that is critical to the success of any solar forecasting task. Additionally, the proposed method does not rely on any exogenous information, such as wind direction and speed, since it infers the spatio-temporal correlation over an area solely from data. A series of verification exercises are conducted to demonstrate the advantages of the proposed method, against carefully selected naïve and state-of-the-art benchmarking methods. Given the high-resolution and adaptive nature, as well as the formal uncertainty quantification, of the forecasts produced, the method and the final datasets herein presented are able to benefit real-time stochastic simulation for smart grid applications at large.

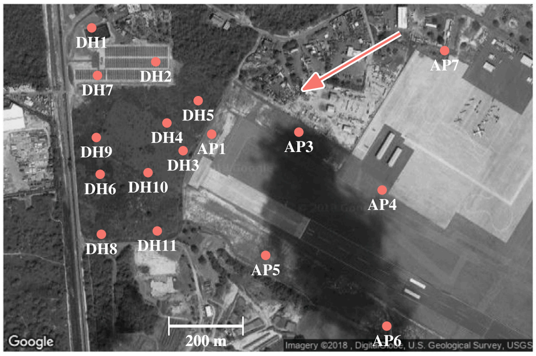


Fig. 1. The geographical arrangement of the 17 stations. The arrow indicates the direction of the prevailing trade winds.

2. Data

High-resolution measurements from the Oahu Solar Measurement Grid (OSMG), installed and maintained by the National Renewable Energy Laboratory (NREL), are considered for empirical demonstration of the proposed method. OSMG consists of 17 stations that are strategically arranged within a 1 km by 1 km area, to study the effects of prevailing trade winds (60° from the North) on solar irradiance variability, see Fig. 1. The project lasted 1 year, from 2010 March through 2011 October. Although this data may appear to be outdated, due to its unique nature, namely, high temporal resolution, high sensor density, and high measurement quality, it remains till today one of (if not) the "go-to" datasets if one is to conduct research in regard to sub-minute irradiance. Indeed, OSMG has been used in a series of forecasting studies [e.g., 26–29], as well as variability studies [e.g., 30–33].

Due to its popularity, the OSMG data has been included in the `SolarData` package in R [34,35], in which appropriate data downloading and quality-control tools are available for a wide range of datasets. Whereas the raw data is provided at a 1-s resolution, they are aggregated into 4-s, 10-s, 30-s, and 1-min resolutions, respectively, to cater for a range of sub-minute applications. In particular, the 4-s resolution aligns with the typical real-time regulation process that is carried out by AGC [36]. The probabilistic forecasts are produced using these aggregated datasets, at all 17 stations. The entire OSMG dataset covers a period from 2010 March through 2011 October. Given its temporal resolution, it would be quite difficult if not impossible to analyze the forecast performance over the entire dataset, because the cloud conditions and weather regimes during this period are diverse. To that end, only 13 days are used for demonstration purposes—they are 2010 July 31, August 1–5, 21, 29, September 5–7, 21, and October 27. In fact, most previous works that considered OSMG used these 13 days [28–32].

Solar irradiance and power time series exhibits yearly and diurnal variations due to the Earth's orbit around the Sun and its self rotation. Since such seasonal cycles can be effectively removed by a clear-sky model and thus result in a local-stationary series, detrending via clear-sky models has become a standard practice in dealing with irradiance or solar power time series [23]. Indeed, failure to do so may result in much higher forecast errors. Following the latest recommendation of a group of 33 solar forecasting experts [13], the McClear model [37] is used here. It is noted that McClear is a physical model, and its smallest resolution is 1 min. Hence, in order to detrend sub-minute irradiance, the values are linearly interpolated. The detrended quantity is known as the clear-sky index, on which the forecasting models are trained and the forecasts are made. Notwithstanding, the forecast clear-sky index is back-transformed to irradiance during verification.

3. Lasso-penalized quantile regression with analog-based preselection

3.1. Challenges and remedies

Suppose the target location where forecasts are generated has $n_s - 1$ neighbors, and for each of these neighbors, n_t lagged variables are considered, then the total number of predictors is $n_s \times n_t$, if the lagged variables at the target location are also considered. For instance, $n_s = 17$ for the Oahu network. This size of n_s alone does not cause problems. The difficulty, however, arises when n_t is large. Under the traditional time series analysis framework (e.g., autoregressive integrated moving average or exponential smoothing), the size of n_t would not be large, since a forecaster usually considers only a few recent lags. Nonetheless, an idea central to weather forecasting is that weather patterns often repeat—patterns that highly resemble each other are known as *weather analogs*, which have been shown useful in solar forecasting [15,38,39]. In this regard, despite some lagged variable being more distant, they can still contribute positively to predictive accuracy—as demonstrated in [40], having $n_t = 100$ substantially improved the root mean square errors at all stations, as compared to the previous setup with $n_t = 10$ [41]. Hence, the first challenge comes from the computation speed requirement of the forecasting model, which is particularly true for online training.

Whenever a large number predictors are present, forecasters can immediately think of dimension reduction techniques. Lasso regression is one of the most popular approaches to date. In principle, it can handle thousands of predictors, as well as the $p > n$ cases, where p is the number of predictors and n is the number of data points [42]. The drawback is however that, under $p > n$, lasso at most selects n variables, which is also a well-known property that often hinders the performance of lasso, see [43,44] for theoretical discussions. On this point, *preselection* becomes a natural remedy, to convert a $p > n$ problem to a $p < n$ problem, without much loss of information, especially when some predictors can be deemed irrelevant. In a solar forecasting context, such preselection usually takes a physical approach, for example, through wind direction and speed [41] or through decorrelation distance [45].

Lastly, even if lasso models could be adequately set up, generating probabilistic predictions, in this case, forecasts, is still difficult. For most regression-based frameworks, such as multiple linear regression (MLR), the prediction interval is usually obtained by estimating the standard error. However, computing the standard error for lasso regression remains challenging [46], and the inventor of lasso, Robert Tibshirani, seems to agree [47]. To estimate the uncertainty associated with a lasso prediction, two options are available, namely, method of dressing and quantile regression. From a methodological viewpoint, method of dressing is a sampling technique, where past errors are sampled and dressed onto the current deterministic-style forecast. On the other hand, quantile regression directly minimizes a set of pinball losses, and thus predicts a set of quantiles, that can be jointly used to quantify the uncertainty. In this paper, the latter is considered.

3.2. Proposed method

In view of the above discussion, this paper proposes a hybrid method that uses lasso-penalized quantile regression with preselection. As the name suggests, it is a two-step method. In the first step, m predictors are selected based on weather analogs, with $m < n_s \times n_t$. Subsequently, lasso-penalized quantile regression (lpQR) is used to fit and forecast the target variable using the preselected predictors. It should be noted that the methods used in both steps are known, however, their combination is novel. Additionally, as shown in the case study below, the proposed method substantially improves the predictive accuracy. In what follows, the basic working principles of both steps are presented.

3.2.1. Analog-based preselection

The analog-based preselection is a similarity-search method. The core idea is to select patterns that are most similar to a length-n *query*—computer scientists refer to these patterns as motifs—from a large historical *database*. The similarity is often defined through Euclidean distance [48]. Denoting forecast timestamp with t, and the variable of interest with Z, the most recent length-n pattern from the target location can be defined:

$$\mathcal{Z}^{(q)} \equiv \{z_{t-1}, z_{t-2}, \dots, z_{t-n}\}, \tag{1}$$

where the superscript (q) denotes query. Given query $\mathcal{Z}^{(q)}$, number of stations n_s, and number of lags n_t, the database then consists of $n_s \times n_t$ patterns defined as:

$$\mathcal{Z}^{(a)}_{jk} \equiv \{z_{t-1-j,k}, z_{t-2-j,k}, \dots, z_{t-n-j,k}\}, \tag{2}$$

with $j = 1, \dots, n_t$ and $k = 1, \dots, n_s$, where $z_{t-1-j,k}$ is the observation taken at time $t-1-j$ at station k, and the superscript (a) denotes analog. By computing the Euclidean distance between $\mathcal{Z}^{(q)}$ and each $\mathcal{Z}^{(a)}_{jk}$, m best analogs can be identified. It should be noted that analog search, though a straightforward task, has hitherto been concerned with computational issues, in that, the exhaustive search is slow by nature. In this regard, several fast-search algorithms have been proposed to remedy the problem [e.g.,49]. In this paper, an ultra-fast search based on k-dimensional tree [50] is considered.

Arranging these m best analogs into a $n \times m$ matrix, that is:

$$\boldsymbol{X} \equiv \begin{pmatrix} z_{t-1-j_1,k_1} & z_{t-1-j_2,k_2} & \cdots & z_{t-1-j_m,k_m} \\ z_{t-2-j_1,k_1} & z_{t-2-j_2,k_2} & \cdots & z_{t-2-j_m,k_m} \\ \vdots & \vdots & \ddots & \vdots \\ z_{t-n-j_1,k_1} & z_{t-n-j_2,k_2} & \cdots & z_{t-n-j_m,k_m} \end{pmatrix}, \tag{3}$$

where j_1 and k_1 denotes the lag and station indexes corresponding to the 1st of the m best analogs, and j_m and k_m denotes the lag and station indexes corresponding to the mth of the m best analogs. At this point, matrix $\boldsymbol{X}$ contains n samples of m predictors, which will be used to predict $\mathcal{Z}^{(q)}$. To simplify the notation, let

$$\boldsymbol{x}_i \equiv \begin{pmatrix} z_{t-i-j_1,k_1} & z_{t-i-j_2,k_2} & \cdots & z_{t-i-j_m,k_m} \end{pmatrix}^\top, \tag{4}$$

with $i = 1, \dots, n$. Similarly, define:

$$y_i \equiv z_{t-i}. \tag{5}$$

At this stage, the predictors and predictand, as would be used subsequently in quantile regression, are ready. To summarize this part, the reader is referred to Fig. 2 for a schematic diagram of the preselection algorithm.

3.2.2. Lasso-penalized quantile regression

To understand quantile regression (QR), it is thought useful to draw the analogy between MLR and QR. In MLR, given n samples of predictor–observation pairs, the equation:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \tag{6}$$

is used to predict the mean of the predictand, where $\boldsymbol{x}_i$ is the ith size-m vector of predictors defined in Eq. (4), and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is a zero-mean homogeneous error. One should note that it is customary to include an intercept term, hence, $\boldsymbol{x}_i$ becomes size-p, where $p = m + 1$. To avoid notation overload, the intercept will not be explicitly written. To estimate the coefficients $\boldsymbol{\beta}$, the sum of squared loss over the training samples is minimized, that is,

$$\hat{\boldsymbol{\beta}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2. \tag{7}$$

With $\hat{\boldsymbol{\beta}}$ and a new predictor vector, namely,

$$\boldsymbol{x}_0 \equiv \begin{pmatrix} z_{t-j_1,k_1} & z_{t-j_2,k_2} & \cdots & z_{t-j_m,k_m} \end{pmatrix}^\top, \tag{8}$$

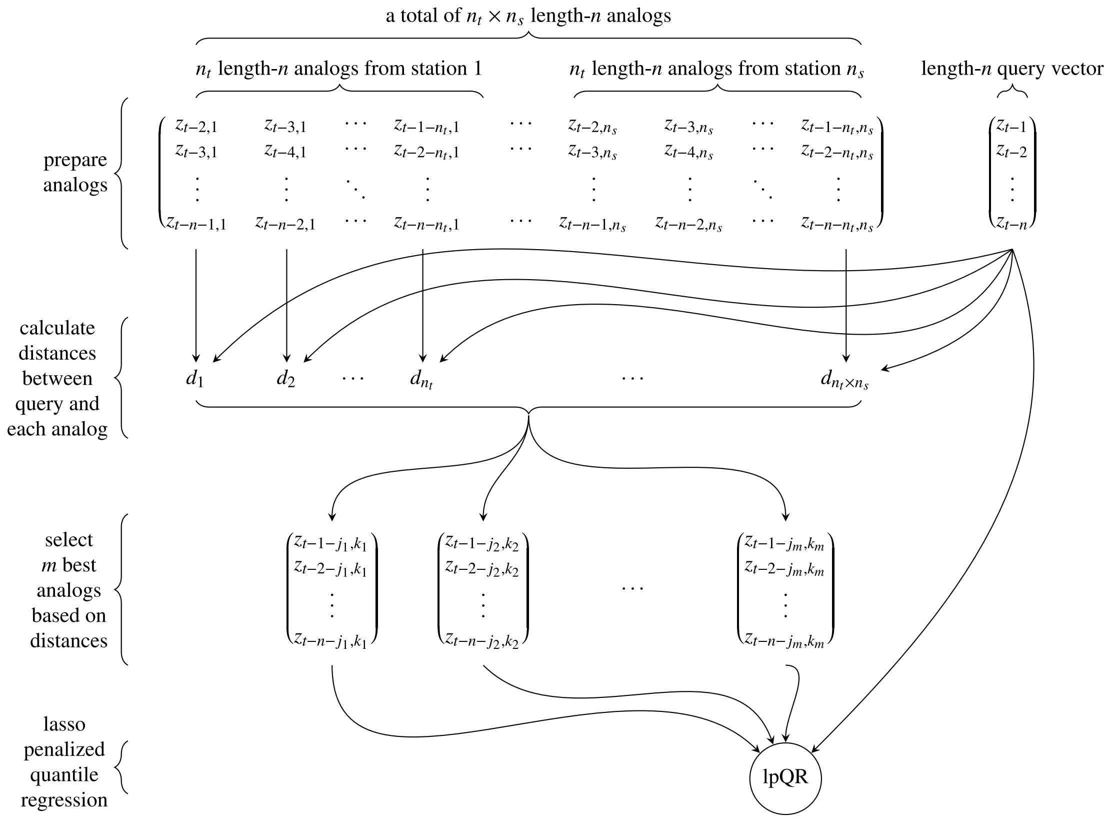

a total of $n_t \times n_s$ length-n analogs
n_t length-n analogs from station 1
n_t length-n analogs from station n_s
length-n query vector
prepare analogs
$z_{t-2,1}$ $z_{t-3,1}$ $\cdots$ $z_{t-1-n_t,1}$ $\cdots$ z_{t-2,n_s} z_{t-3,n_s} $\cdots$ z_{t-1-n_t,n_s}
$z_{t-3,1}$ $z_{t-4,1}$ $\cdots$ $z_{t-2-n_t,1}$ $\cdots$ z_{t-3,n_s} z_{t-4,n_s} $\cdots$ z_{t-2-n_t,n_s}
$z_{t-n-1,1}$ $z_{t-n-2,1}$ $\cdots$ $z_{t-n-n_t,1}$ $\cdots$ z_{t-n-1,n_s} z_{t-n-2,n_s} $\cdots$ z_{t-n-n_t,n_s}
z_{t-1} z_{t-2} z_{t-n}
calculate distances between query and each analog
d_1 d_2 $\cdots$ d_{n_t} $\cdots$ $d_{n_t \times n_s}$
select m best analogs based on distances
z_{t-1-j_1,k_1} z_{t-2-j_1,k_1} z_{t-n-j_1,k_1}
z_{t-1-j_2,k_2} z_{t-2-j_2,k_2} z_{t-n-j_2,k_2}
z_{t-1-j_m,k_m} z_{t-2-j_m,k_m} z_{t-n-j_m,k_m}
lasso penalized quantile regression
lpQR

Fig. 2. Schematic diagram of the proposed algorithm. The algorithm starts by define a total of $n_t \times n_s$ analogs using lagged variables; the distance between the query (the latest data from the station of interest) and each analog is then calculated; m best analogs are selected based on the smallest distances; and the selected analogs are used as predictors for the lasso-penalized quantile regression (lpQR), which generates the forecast.

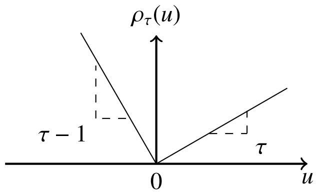


Fig. 3. Pinball loss.

the mean of $\hat{y}_0$ is $\mathbb{E}(\hat{y}_0) = \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}$. By the definition in Eq. (5), y_0 is just z_t, i.e., the materialized value of the forecast variable Z at the forecast time t.

On the other hand, since QR predicts quantiles, instead of minimizing the squared loss in Eq. (7), it minimizes the pinball loss, $\rho_\tau(\cdot)$, i.e.,

$$\hat{\boldsymbol{\beta}}(\tau) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right), \tag{9}$$

where

$$\rho_\tau(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) = \begin{cases} \tau |y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}| & \text{for } \boldsymbol{x}_i^\top \boldsymbol{\beta} \le y_i, \\ (1-\tau)|y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}| & \text{for } \boldsymbol{x}_i^\top \boldsymbol{\beta} \ge y_i, \end{cases} \tag{10}$$

where $\tau \in (0, 1)$ is used to index the τth quantile. A graphical illustration of the pinball loss is given in Fig. 3.

The analogy between lasso and QR-lasso is similar. In that, the MLR model parameter estimated under lasso penalty is expressed as:

$$\hat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1, \tag{11}$$

where the second term shrinks the unconstrained least-squares estimator towards $\boldsymbol{\beta}_0$, which is typically set to zero, and λ is known as the penalty strength. The lpQR estimator,

$$\hat{\boldsymbol{\beta}}(\tau, \lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right) + \lambda \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1, \tag{12}$$

is again written by replacing the squared loss in Eq. (11) with pinball loss. Once $\hat{\boldsymbol{\beta}}(\tau, \lambda)$ is estimated, with a new predictor vector, denoted using $\boldsymbol{x}_0$, the corresponding τth quantile, or $q_{\tau,0}$, can be predicted using:

$$\hat{q}_{\tau,0} = \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}(\tau, \lambda). \tag{13}$$

As evidence by Eq. (12), if a forecaster is interested in a collection of $\tau \in \mathbb{Q}$ quantiles, Q regressions need to be fitted, where Q is the cardinality of the set $\mathbb{Q}$. On this point, the efficiency of the lpQR routine becomes important. Fortunately, QR and its variants have been extensively studied in the literature, and the reader is referred to [51] for most, if not all, of its technical details. In this paper, the lpQR routine in the `quantreg` package in R is used without loss of generality. A set of $Q = 21$ quantiles, namely, $\tau \in \{0.025, 0.05, 0.1, \dots, 0.9, 0.95, 0.975\}$, is used throughout the paper. In other words, aside from the 0.025 and 0.975 quantiles, all other quantiles are spaced evenly at a distance of 0.05.

4. Benchmarks and evaluation metrics

A total of five benchmarking methods are used to gauge the performance of the proposed method. In weather forecasting, climatology and persistence are two most popular reference methods that are used in almost every occasion [52]. However, both climatology and persistence are commonly defined for deterministic forecasts. To apply them in probabilistic forecasting, some adjustments need to be made.

Climatology refers to the long-term behavior of weather. In deterministic forecasting, the most common form of climatology is to issue the mean of the variable of interest as forecasts. Moreover, for simplicity, the mean is often computed from the verification samples [52]. Analogous to the deterministic case, climatology for the probabilistic case is the unconditional distribution of the variable of interest [53,54]. That is, given a forecast variable Z, its distribution, $F_Z(z)$, is used as the predictive distribution, regardless of the forecasting situation and forecast horizon. When $F_Z(z)$ is unknown, the empirical cumulative distribution function (ECDF) is used instead, denoted by $\hat{F}_Z(z)$. Since climatology issues the same forecast for every forecasting situation, it lacks discrimination, which refers to the ability to discriminate different forecast situations as to issue different forecasts. However, it is perfectly calibrated by construct, which refers to the statistical consistency between the probabilistic forecasts and observations. In this paper, an underlined small-caps word is used to denote a method, e.g., CLIM refers to the climatology benchmark hereafter.

Persistence takes the most recent observation as forecasts. For one-step-ahead forecasting, the forecast for time t is the observation made at time $t-1$. To extend persistence to the probabilistic case, forecasters often opt for the persistence ensemble (PeEn). PeEn uses m most recent observations to form an ECDF, which is then used as the predictive distribution. Here, m is set to 21, so that the observations can be assumed to represent the Q quantiles. This method is denoted as TMP-PEEN, where “Tmp” stands for “temporal”. Given the fact that the PeEn forecasts can also be constructed using recent observations made by spatial neighbors, another benchmark used in this paper is SPT-PEEN, where “Spt” stands for “spatial”. For this method, the most recent observations from the n_s stations are gathered and form an ECDF. Subsequently, based on that ECDF, $\tau \in \mathbb{Q}$ quantiles are drawn. In clear contrast to climatology, PeEn, due to its small sample size, is usually sharp, which refers to the fact that the ensemble spread is small. Nonetheless, it lacks calibration.

Recall that the proposed method first preselects m analogs. If letting $m = Q$, $\boldsymbol{x}_0$ in Eq. (8) is in fact the analog ensemble (AnEn) forecast [38,50]. As discussed earlier in Section 3.1, AnEn is a popular method for weather forecasting, especially when dynamical ensemble forecasts with perturbed initial conditions are unavailable [55]. In other words, AnEn can be viewed as a post-processing method, with which the forecaster is able to convert a set of deterministic forecasts to a probabilistic one [55]. To that end, ANEN is used as the fourth benchmark here. It follows that the proposed method is in fact a hybrid model with AnEn and lpQR—it is therefore denoted as ANEN+LPQR hereafter.

The last benchmark resembles ANEN+LPQR. Instead of using the preselected predictors for lpQR, the benchmark takes only the lag-1 variables from the n_s locations as predictors. This setup is similar to a vector autoregressive model of order 1 (VAR1). Whereas VAR1 predicts the mean, this benchmark predicts the τth quantile. This benchmarking method is denoted as LAG1+LPQR, namely, the lasso-penalized quantile regression with lag-1 predictors. Theoretically, one can also include more lagged variables into modeling, e.g., those from lag-2, -3, up to n_t. However, this results in $n_s \times n_t$ predictors, where computation speed becomes an issue—this in fact constitutes our initial motivation of performing the preselection. Hence, this benchmark restricts the lagged variables to order 1.

It should be clearly noted that, for the reason given in Section 2, *all* proposed and benchmarking models are trained using clear-sky index, but their forecasts are back-transformed to irradiance for verification. To verify the probabilistic forecast performance, four metrics are used, namely, prediction interval coverage probability (PICP), prediction interval average width (PIAW), continuous ranked probability score (CRPS), and the pinball loss. PICP is a measure for calibration, for calibrated forecasts, PICP should be close to the nominal coverage probability, i.e., $1-\alpha$, with $\alpha = 0.05$ used throughout the paper. PIAW is simply the average width of the $1-\alpha$ central prediction interval. CRPS is a strictly proper scoring rule, that is advised to be used in most, if not all, probabilistic forecasting applications [56], with no exception for solar [57]. Lastly, the pinball loss, also known as the tick loss or check loss, is a highly popular choice in probabilistic load forecasting [58].

5. Results

In this section, the proposed method is applied to the aforementioned OSMG dataset. Whereas it is clear that $n_s = 17$, the choice of n_t needs to be made. It was previously reported that the average wind speed over the selected 13 days is about 10 m/s [30], given the geographical area of the network—furthest stations are about 1 km apart—it implies that a cloud would take, on average, 100 s to propagate from one end of the network to the other. To capture the movement of slower clouds, n_t value is set according to three times of the average case, that is, 300 s. In other words, for 4-s-, 10-s-, 30-s- and 1-min-averaged data, their n_t values are 75, 30, 10, and 5, respectively.

On the other hand, the length of query, n, which is also the training data length for each forecast, is set to be 150 for all cases, to ensure sufficient samples for QR regression [40]. For each query, the number of analogs m is set to 21, which is a standard choice in solar forecasting [38,59]. Furthermore, a zenith angle filter of $z < 80°$ is used, beyond which the irradiance values are set to 0. This is because even the best clear-sky models at low-sun conditions are not reliable [23]. Another practical issue is that forecasts for the first $n + n_t$ daylight timestamps in each day cannot be issued, since the training periods required for those forecasts overlap with nighttime (or $z > 80°$). At the moment, there is no appealing solution that can address this issue, hence forecasting at early mornings is deferred to future works.

The experiment which we wish to conduct spans three dimensions: (1) resolution of data, (2) day number, and (3) station number. In other words, the forecasting can be performed at each resolution, each day, and each station. On one hand, if we are to lump the forecast performance calculation across all dimensions, it would hinder the interpretation of the results. It is, therefore, thought appropriate to present the results in a more progressive manner. More specifically, Section 5.1 shows the results for 1 resolution, 1 day, and 1 station; Section 5.2 shows the results for 1 resolution, 1 day, and all stations; Section 5.3 shows the results for 1 resolution, all days, and all stations; and finally, Section 5.4 shows the results for all resolutions, 1 day, and all stations.

5.1. Performance for 1 resolution, 1 day, and 1 station

To exemplify the procedure, forecasts made for 1 station, at 1 time resolution, and over 1 day, are first discussed. Without loss of generality, 1-min-ahead forecasts at DH3 station on 2010 July 31 are used here. After applying the zenith angle filter, a total of 689 timestamps remain; a snapshot is shown in Fig. 4, with the transient of DH3 shown in blue, and that of the remaining stations shown in gray. It can be seen that some gray lines lead the DH3 time series, whereas others lag behind it; this justifies the use of spatial neighbors during forecasting, since the leading (up-wind) stations are able to provide information as to the probable ramps that may occur at the focal station at a later instance. Recall that the first $n+n_t$ (155 in this case) forecasts cannot be issued due to insufficient training samples, the number of verification samples is hence 534. For each of these 534 1-min timestamps, 21 ANEN+LPQR models are trained, one for each $\tau \in \mathbb{Q}$ quantile. What this implies is that rolling forecast is performed, and the model is re-trained after every minute.

Table 1 shows the performance of ANEN+LPQR alongside five benchmarks, in terms of four evaluation metrics. It can be seen that ANEN+LPQR has a PICP of 93.3%, which is close to the nominal coverage of 95%. Aside from climatology, which is calibrated by construct—it has a PICP of 95.3%, which confirms this fact—all other benchmarks fall short of the nominal coverage. On the other hand, ANEN+LPQR has

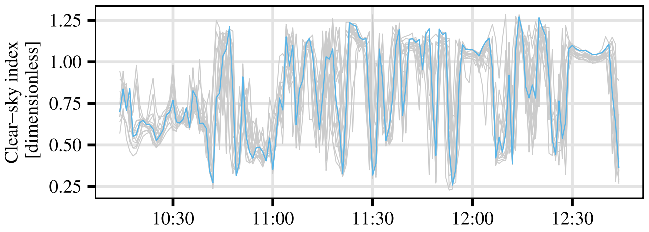


Fig. 4. A query window (150 data points) of 1-min clear-sky index time series on 2010 July 31. DH3 is in blue, and its neighbors are in gray.

Table 1
Forecast performance at DH3 on 2010 July 31 using 1-min data. Besides PICP which is in %, all other metrics are in W/m^2.

Model	PICP	PIAW	CRPS	Pinball
Clim	95.3	817.2	137.3	66.0
Tmp-PeEn	87.8	627.8	120.2	58.4
Spt-PeEn	72.8	322.9	84.7	41.5
AnEn	82.2	412.8	87.5	42.8
Lag1+lpQR	91.6	489.2	63.9	31.3
AnEn+lpQR	93.3	482.7	54.9	27.3

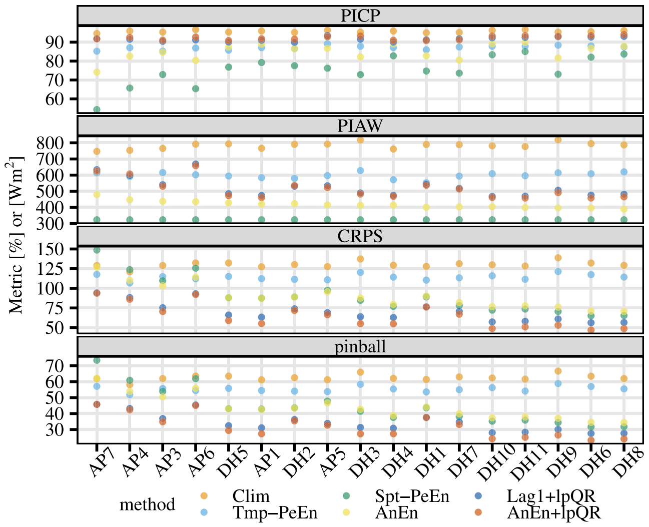


Fig. 5. Performance of 1-min-ahead forecasts on 2010 July 31. The stations are arranged in the along-wind direction. Colorbind-friendly palette is used.

a PIAW of 482.7 W/m^2, which is the third sharpest, after Spt-PeEn and AnEn. Since the overarching rule for making good probabilistic forecasts is to minimize sharpness subject to calibration [60], there is often a trade-off between PICP and PIAW. In this regard, CRPS and pinball loss as composite scores reflect the overall quality of the forecast quantiles—AnEn+lpQR has the best scores in terms of these, followed by Lag1+lpQR. Particularly interesting is that AnEn+lpQR is better than Lag1+lpQR in terms of both PICP and PIAW, which suggests that the proposed method is more reliable and sharper than Lag1+lpQR.

The results shown in Table 1 can be further interpreted as follows. Firstly, in terms of the composite scores, that is, CRPS and pinball loss, Clim and Tmp+PeEn are largely unsatisfactory. This is because both benchmarking methods do not consider any spatial information during forecasting. As clouds are the primary factor influencing surface solar radiation, whether or not the forecasting model has the capacity to capture the cloud movement, as also highlighted by Yang [12], has hitherto been viewed as the principal consideration in evaluating whether or not a solar forecasting work can be taken as state-of-the-art. Among the remaining four methods, it is evident that those with an additional lpQR step, i.e., AnEn+lpQR and Lag1+lpQR, outperform those without, i.e., Spt+PeEn and AnEn. Since lpQR assigns different weights to different predictors, the result reveals that further optimization of predictors through weighting is necessary and beneficial. Lastly, it is worth-mentioning that, although the number of predictors as used by AnEn+lpQR is comparable to that used by Lag1+lpQR, the former has an advantage over the latter, owing to the fact that the predictors of AnEn+lpQR are more comprehensive in describing the temporal correspondence among the stations. This, in the main, confirms the need for the preselection step during spatio-temporal forecasting.

5.2. Performance for 1 resolution, 1 day, and all stations

Extending the case study in Section 5.1 to the 1-resolution, 1-day, and all-station case, the same procedure is repeated using 1-min data from 2010 July 31 for each station. Fig. 5 shows the forecast performance in terms of the four metrics. To facilitate visualization, the stations are ordered in the along-wind direction, see Fig. 1. More specifically, AP7 without any preceding station is placed first, whereas DH8 with 16 preceding stations is placed last. Naturally, the stations with more up-wind stations tend to have a higher accuracy, owing to the availability of analogs that lead the queries. In contrast, stations without up-wind stations (cf. Fig. 1), namely, AP7, AP4, AP6, DH5, DH2, and DH1 perform poorly, because they have no predictor that can provide useful information regarding the incoming clouds.

With little surprise, all previous observations and analyses made on the 1-station case also apply in the all-station case. In short, AnEn+lpQR has the best performance in terms of CRPS and pinball loss at all stations. This provides empirical evidence to the general validity of the preselection algorithm. Similarly, Clim again has the best PICP (closest to 95%) at all stations, followed by AnEn+lpQR. On the other hand, AnEn and Spt-PeEn obtained low PIAW (i.e., sharp predictive distributions), at a cost of low PICP (i.e., insufficient probability coverage). Generally speaking, sharpness depends on the diversity in the ensemble. If analogs are few or with high similarity, a low PIAW is often observed. In that, it can be said that AnEn and Spt-PeEn lack diversity.

5.3. Performance for 1 resolution, all days, and all stations

Extending the case study in Section 5.2 to the 1-resolution, all-day, and all-station case, the forecasts are generated using 1-min data for each of the 13 days and for each station. The error metrics for each method, on each day, and at each station are plotted in Fig. 6. In the top left subplot, the color bar corresponds to the difference between the PICP and the nominal coverage probability. In that subplot, a dark tile indicates that the forecasts for that station, that day are calibrated. The other three subplots use different colorblind-friendly palettes; since all of these metrics are negatively oriented (i.e., the smaller the better), the lighter the color of the tile is, the better the result is.

From Fig. 6, several remarks can be made. In general, the performance of Clim and Tmp-PeEn resembles each other, and is lower than that of other methods. This can be directly attributed to the high variability of 1-min clear-sky index, as shown in Fig. 4. In such cases, the predictive ability of temporal-only methods is limited, since one cannot identify incoming clouds solely based on the past irradiance transient at the target station. It follows that in order to improve forecast accuracy, spatial information must be utilized, as evidenced by the better performance of Spt-PeEn and AnEn. Both Spt-PeEn and AnEn directly take past measurements from neighboring stations as forecasts. Although some level of optimization is involved during the process—recall that the analogs are searched based on smallest Euclidean distances—without post-processing, the predictive power is still

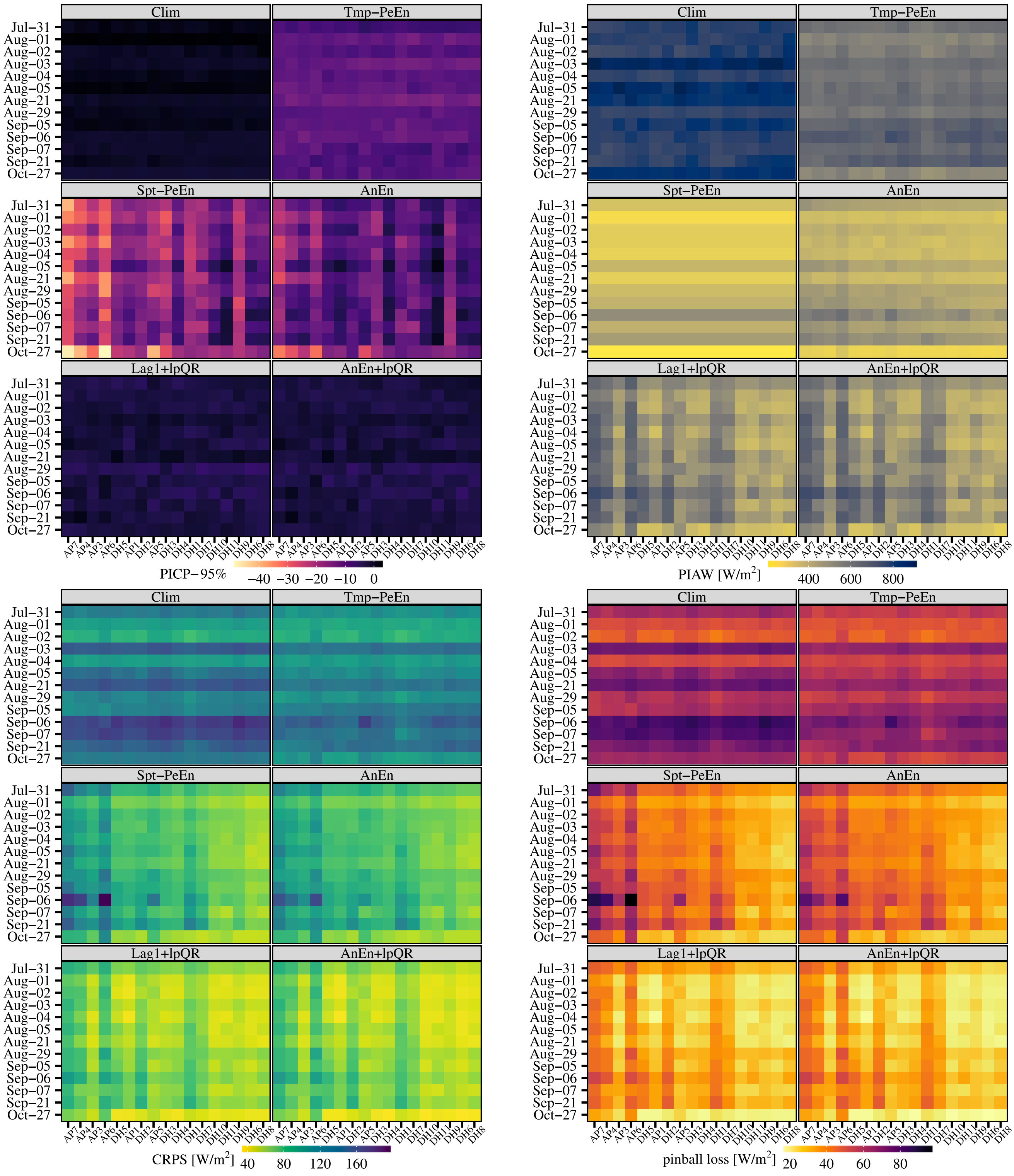


Fig. 6. PICP (top left), PIAW (top right), CRPS (bottom left) and pinball loss (bottom right) of 1-min-ahead forecasts over 13 days at all stations. Colorblind-friendly palettes are used.

limited. To that end, the lpQR-based methods, namely, Lag1+lpQR and AnEn+lpQR, that post-process the analog forecasts, are clearly more advantageous.

At this stage, it is useful to introduce the concept of skill score. In meteorological forecasting, skill score refers to the relative improvement made by forecasts of interest on top of some naïve reference forecasts [52]. Denoting the observations, forecasts of interest, and reference forecasts by x, f, and r, respectively, the skill score, s, is

$$s = 1 - \frac{A(x, f)}{A(x, r)}, \tag{14}$$

where A is some accuracy measure, and s is often expressed in percentage. In that, if f is no better than r, $s \leq 0$, otherwise, $s > 0$, which marks the minimal level of acceptance to use f instead of r. A rule-of-thumb for selecting the standard of reference in skill score

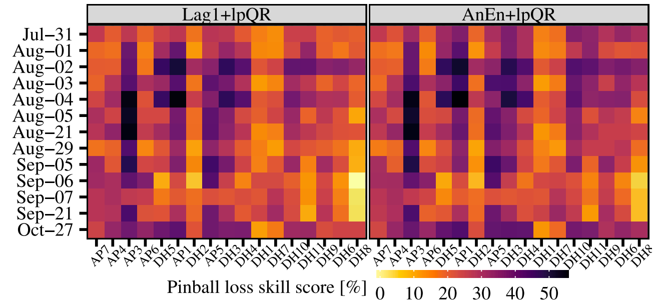


Fig. 7. Pinball loss skill scores of 1-min-ahead forecasts over 13 days at all stations. AnEn is used as the reference method.

Table 2
Averaged forecast performance over all stations, using 4-s, 10-s, 30-s, and 1-min data, on 2010 July 31. Only Lag1+lpQR and AnEn+lpQR results are reported with the rest omitted.

Model	PICP	PIAW	CRPS	Pinball
Lag1+lpQR 4-s	89.7	261.1	26.7	13.5
AnEn+lpQR 4-s	88.8	239.1	24.7	12.6
Lag1+lpQR 10-s	91.1	352.6	39.4	19.7
AnEn+lpQR 10-s	90.9	335.1	37.6	18.9
Lag1+lpQR 30-s	91.6	426.7	51.6	25.6
AnEn+lpQR 30-s	91.6	418.1	49.2	24.6
Lag1+lpQR 1-min	91.6	522.8	69.7	34.2
AnEn+lpQR 1-min	92.3	514.6	64.5	31.9

computation is that r has to be naïve, but at the same time, possesses a decent level of accuracy, so that s can be used to compare forecasts—with appropriate caveats—made by different methods, over different time periods, and at different locations [54,61]. For that reason, AnEn is used as the reference method, to gauge the relative performance of Lag1+lpQR and AnEn+lpQR. Without loss of generality, the pinball loss is used as the accuracy measure A.

Fig. 7 depicts the pinball loss skill scores of Lag1+lpQR and AnEn+lpQR with respect to AnEn—the negative scores for other methods are omitted. It can be seen that forecasts using the lpQR-based methods produce positive skill scores for all days and all locations. On average, Lag1+lpQR has a skill score of 26.02%, whereas the average skill score of AnEn+lpQR is 27.80%. Considering the fact that AnEn itself is already quite a high-performance probabilistic solar forecasting method—it has been shown to be the best reference method among several popular benchmarks [59]—the skill scores attained by AnEn+lpQR are remarkable.

5.4. *Performance for all resolutions, 1 day, and all stations*

Lastly, the same exercise in Section 5.2 is repeated using data with different resolutions (4-s, 10-s, 30-s, and 1-min). Here, to save space and computational effort, only results from a single day, namely, 2010 July 31, are reported. One should note that although only a day's results are shown, the number of verification data points are in fact abundant. More specifically, at each station, the numbers of verification samples are 10103, 3472, 1217 and 534, for the four resolutions respectively. Additionally, since these 13 days have the same weather features where trade-wind-carried broken clouds are predominant, the results for 13 days would resemble highly that for a single day, as also evidenced from the previous case studies. Nevertheless, readers can explore the remaining results on their own, see Section 6. Table 2 shows the results. AnEn+lpQR outperforms Lag1+lpQR under PIAW, CRPS, pinball loss for all data resolutions. Evidently, the predictive distributions of AnEn+lpQR has higher quality than that of Lag1+lpQR.

6. Uptake and application

6.1. *Data, code, and uptake of the present result*

As mentioned in the introduction, all code and final forecast datasets will be released on Github, see https://github.com/dazhiyang/AnEnlpQR. However, owing to the NREL data policy, the present authors cannot directly share the raw data.[1] However, downloading the raw data can be done with a few clicks from the NREL Measurement and Instrumentation Data Center (MIDC), at https://midcdmz.nrel.gov/. Once the raw daily data files are downloaded and unzipped, the user could process and aggregate the raw files using the `Arrange.R` script. This script will automatically generated the input files needed to produce forecasts. Subsequently, the user can run the `GenFcst.R` script, which will produce the output files that contain the final forecast datasets. At the moment, the script produces forecasts for one day, one aggregation interval, at a time. Lastly, to reproduce the results reported in this paper, four R scripts named after the section title, namely, `Section V A.R` to `Section V D.R` are also provided. All R scripts include extensive comments that can help a user to understand the forecasting workflow, as well as various settings used in each forecasting method.

Clearly, these datasets have many applications. For instance, using the well-known PV modeling packages, such as `pvlib` packages in Python or NREL's System Advisor Model, one can convert these irradiance forecasts to PV power forecasts by supplying user-defined PV system specifications. Henceforth, using these PV power forecasts that carry the notion of uncertainty, one can perform real-time stochastic simulation of all sorts, by drawing samples from the predictive distributions sequentially.

6.2. *Applicability*

Although the uptake of the current results has been described, one must note the fact that the OSMG dataset is by no means one that is commonly available. High-resolution, high-quality irradiance monitoring networks are rare at the moment. Hence, there is a need to discuss issues pertaining to the applicability and transferability of the method proposed in this paper. More specifically, the discussion is separated into two parts. One of those is on the monitoring network design, and the other on the utilization of distributed PV systems as sensors.

The forecast horizon supported by a sensor network is closely tied to the network's spatial coverage, sampling rate, and number of sensors within the network. In that, monitoring network design becomes relevant. Generally speaking, before designing a network, one must understand the local meteorological conditions, which can be done by analyzing satellite-derived or reanalysis data, particularly those depicting long-term, representative wind information. Firstly, it is known *a priori* that allocating more sensors in up-wind is necessary. To make sure incoming clouds from all directions can be detected, Chen et al. [62] designed a network which is arranged in concentric circles around the PV plant of interest, such that the outer and inner radiometer pairs can act as cloud direction sensors. Secondly, the station spacing must be designed according to the average *in situ* wind speed, such that the movement of clouds can be captured by the sensor network [62,63]. If the stations are placed too close to one another, clouds may pass the target location during unit time, i.e., the step length of forecast. On the other hand, if the stations are too far way from each other, clouds may have deformed or dissipated when they reach the target location. Besides the consideration given to wind direction and speed, also important is network re-design, which deals with optimally adding, relocating, or removing sensors from an existing network [see 64,65, for example procedures].

[1] The reader may contact the corresponding author for arranged data.

Setting up a sensor network may be costly. Hence, the economic viability must be justified, by considering the value-add of forecasts to the plant operation. Whereas such cost issue may not present a problem to large utility-scale, multi-MW-size PV plant owners, smaller residential PV system owners are unlikely to construct their own sensor networks. There is, however, a work-around, that is, using distributed PV system as sensors. Such possibility has been studied by Elsinga and van Sark [66] and Lonij et al. [67], among others. Considering the distancing of residential houses of a typical American town, the forecasting method and its associated horizon may apply just as well. In this case, instead of training the lasso-penalized quantile regression on clear-sky index, one may train the model on normalized PV power, or clear-sky index for PV [68].

7. Conclusion

This paper is, in the main, concerned with generating sub-minute solar irradiance forecasts, for high-temporal-resolution stochastic simulation, as required by those studies on secondary control in power systems. Similarly, such high-resolution forecasts can also be deemed useful when investigating hybrid energy system management strategies, with which applications, such as battery charging/discharging under variable solar generation, can be optimized. Indeed, one of the major barriers preventing power system engineers and solar engineers from using realistic forecasts in stochastic simulation studies is the unavailability of readily available forecast dataset, this paper presents such an opportunity by issuing state-of-the-art, spatio-temporal, probabilistic, online-trained, high-resolution, high-accuracy forecasts that are publicly available.

Technique-wise, the proposed method uses the lasso-penalized quantile regression with preselection to perform online forecasting of sub-minute solar irradiance over a 1 km by 1 km area. Through extensive validation against widely recognized benchmarks and previous published forecasting frameworks, the superiority of the method can be confirmed at once. More specifically, forecasts generated by the proposed method is able to achieve a pinball loss skill score up to 55%, with analog ensemble being the reference method. This high forecast skill can be chiefly attributed to the preselection of spatio-temporal predictor, which chooses only the most relevant information as to the forecasting of the target variable. In terms of computational efficiency, the preselection algorithm is highly competitive, in that, one of the world's fastest similarity-search routines is used to power the analog matching. That said, another contribution, and arguably the most important one, is that probabilistic method is applied on sub-minute irradiance forecasting for the first time. In fact, as the field of energy meteorology has become increasingly concerned with uncertainty quantification, probabilistic method would no doubt be regarded as the default moving forward.

CRediT authorship contribution statement

Dazhi Yang: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Gokhan Mert Yagli:** Methodology, Validation, Writing – review & editing. **Dipti Srinivasan:** Validation, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Ran X, Miao S, Jiang Z, Xu H. A framework for uncertainty quantification and economic dispatch model with wind–solar energy. Int J Electr Power Energy Syst 2015;73:23–33. http://dx.doi.org/10.1016/j.ijepes.2015.03.023, URL: http://www.sciencedirect.com/science/article/pii/S0142061515001581.

[2] Carreño IL, Ramakrishna R, Scaglione A, Arnold D, Roberts C, Ngo S-T, Peisert S, Pinney D. Soda: An irradiance-based synthetic solar data generation tool. In: 2020 ieee international conference on communications, control, and computing technologies for smart grids (smartgridcomm). 2020, p. 1–6. http://dx.doi.org/10.1109/SmartGridComm47815.2020.9302941.

[3] Ganger D, Zhang J, Vittal V. Forecast-based anticipatory frequency control in power systems. IEEE Trans Power Syst 2018;33(1):1004–12. http://dx.doi.org/10.1109/TPWRS.2017.2705761.

[4] Tao Z, Moncada JA, Poncelet K, Delarue E. Review and analysis of investment decision making algorithms in long-term agent-based electric power system simulation models. Renew Sustain Energy Rev 2021;136:110405. http://dx.doi.org/10.1016/j.rser.2020.110405, URL: http://www.sciencedirect.com/science/article/pii/S1364032120306936.

[5] Kharrazi A, Sreeram V, Mishra Y. Assessment techniques of the impact of grid-tied rooftop photovoltaic generation on the power quality of low voltage distribution network - A review. Renew Sustain Energy Rev 2020;120:109643. http://dx.doi.org/10.1016/j.rser.2019.109643, URL: http://www.sciencedirect.com/science/article/pii/S1364032119308500.

[6] Talari S, Shafie-khah M, Osário GJ, Aghaei J, ao P.S. Catalão J. Stochastic modelling of renewable energy sources from operators' point-of-view: A survey. Renew Sustain Energy Rev 2018;81:1953–65. http://dx.doi.org/10.1016/j.rser.2017.06.006, URL: http://www.sciencedirect.com/science/article/pii/S1364032117309437.

[7] Sharafi M, ElMekkawy TY. Stochastic optimization of hybrid renewable energy systems using sampling average method. Renew Sustain Energy Rev 2015;52:1668–79. http://dx.doi.org/10.1016/j.rser.2015.08.010, URL: http://www.sciencedirect.com/science/article/pii/S1364032115008527.

[8] Scolari E, Reyes-Chamorro L, Sossan F, Paolone M. A comprehensive assessment of the short-term uncertainty of grid-connected PV systems. IEEE Trans Sustain Energy 2018;9(3):1458–67. http://dx.doi.org/10.1109/TSTE.2018.2789937.

[9] Chen X, Du Y, Wen H, Jiang L, Xiao W. Forecasting-based power ramp-rate control strategies for utility-scale PV systems. IEEE Trans Ind Electron 2019;66(3):1862–71. http://dx.doi.org/10.1109/TIE.2018.2840490.

[10] Zheng Y, Zhao J, Song Y, Luo F, Meng K, Qiu J, Hill DJ. Optimal operation of battery energy storage system considering distribution system uncertainty. IEEE Trans Sustain Energy 2018;9(3):1051–60.

[11] Yang D, Kleissl J, Gueymard CA, Pedro HTC, Coimbra CFM. History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining. Sol Energy 2018;168:60–101. http://dx.doi.org/10.1016/j.solener.2017.11.023, URL: https://www.sciencedirect.com/science/article/pii/S0038092X17310022.

[12] Yang D. A guideline to solar forecasting research practice: Reproducible, operational, probabilistic or physically-based, ensemble, and skill (ROPES). J Renew Sustain Energy 2019;11(2):022701. http://dx.doi.org/10.1063/1.5087462.

[13] Yang D, Alessandrini S, Antonanzas J, Antonanzas-Torres F, Badescu V, Beyer HG, Blaga R, Boland J, Bright JM, Coimbra CF, David M, Frimane Â, Gueymard CA, Hong T, Kay MJ, Killinger S, Kleissl J, Lauret P, Lorenz E, van der Meer D, Paulescu M, Perez R, Perpiñán-Lamigueiro O, Peters IM, Reikard G, Renné D, Saint-Drenan Y-M, Shuai Y, Urraca R, Verbois H, Vignola F, Voyant C, Zhang J. Verification of deterministic solar forecasts. Sol Energy 2020;210:20–37. http://dx.doi.org/10.1016/j.solener.2020.04.019, URL: http://www.sciencedirect.com/science/article/pii/S0038092X20303947 Special Issue on Grid Integration.

[14] Mayer MJ, Gróf G. Extensive comparison of physical models for photovoltaic power forecasting. Appl Energy 2021;283:116239. http://dx.doi.org/10.1016/j.apenergy.2020.116239, URL: https://www.sciencedirect.com/science/article/pii/S0306261920316330.

[15] Zhang X, Li Y, Lu S, Hamann HF, Hodge B, Lehman B. A solar time based analog ensemble method for regional solar power forecasting. IEEE Trans Sustain Energy 2019;10(1):268–79. http://dx.doi.org/10.1109/TSTE.2018.2832634.

[16] Carriere T, Vernay C, Pitaval S, Kariniotakis G. A novel approach for seamless probabilistic photovoltaic power forecasting covering multiple time frames. IEEE Trans Smart Grid 2020;11(3):2281–92. http://dx.doi.org/10.1109/TSG.2019.2951288.

[17] Allen MR, Stainforth DA. Towards objective probabilistic climate forecasting. Nature 2002;419(6903):228.

[18] Gneiting T, Katzfuss M. Probabilistic forecasting. Annu Rev Stat Appl 2014;1(1):125–51. http://dx.doi.org/10.1146/annurev-statistics-062713-085831.

[19] Appino RR, González Ordiano JA, Mikut R, Faulwasser T, Hagenmeyer V. On the use of probabilistic forecasts in scheduling of renewable energy sources coupled to storages. Appl Energy 2018;210:1207–18. http://dx.doi.org/10.1016/j.apenergy.2017.08.133, URL: https://www.sciencedirect.com/science/article/pii/S0306261917311492.

[20] Li B, Zhang J. A review on the integration of probabilistic solar forecasting in power systems. Sol Energy 2020;210:68–86. http://dx.doi.org/10.1016/j.solener.2020.07.066, URL: https://www.sciencedirect.com/science/article/pii/S0038092X20307982 Special Issue on Grid Integration.

[21] Agoua X, Girard R, Kariniotakis G. Probabilistic models for spatio-temporal photovoltaic power forecasting. IEEE Trans Sustain Energy 2019;10(2):780–9. http://dx.doi.org/10.1109/TSTE.2018.2847558.

[22] Sanjari MJ, Gooi HB. Probabilistic forecast of PV power generation based on higher order Markov chain. IEEE Trans Power Syst 2017;32(4):2942–52. http://dx.doi.org/10.1109/TPWRS.2016.2616902.

[23] Yang D. Choice of clear-sky model in solar forecasting. J Renew Sustain Energy 2020;12(2):026101. http://dx.doi.org/10.1063/5.0003495.

[24] Urraca R, Gracia-Amillo AM, Huld T, de Pison FJM, Trentmann J, Lindfors AV, Riihelä A, Sanz-Garcia A. Quality control of global solar radiation data with satellite-based products. Sol Energy 2017;158:49–62. http://dx.doi.org/10.1016/j.solener.2017.09.032, URL: https://www.sciencedirect.com/science/article/pii/S0038092X17308046.

[25] Killinger S, Engerer N, Müller B. QCPV: A quality control algorithm for distributed photovoltaic array power output. Sol Energy 2017;143:120–31. http://dx.doi.org/10.1016/j.solener.2016.12.053, URL: https://www.sciencedirect.com/science/article/pii/S0038092X16306600.

[26] van der Meer D, Yang D, Widén J, Munkhammar J. Clear-sky index space-time trajectories from probabilistic solar forecasts: Comparing promising copulas. J Renew Sustain Energy 2020;12(2):026102. http://dx.doi.org/10.1063/1.5140604.

[27] Amaro e Silva R, Haupt SE, Brito MC. A regime-based approach for integrating wind information in spatio-temporal solar forecasting models. J Renew Sustain Energy 2019;11(5):056102. http://dx.doi.org/10.1063/1.5098763.

[28] Aryaputera AW, Yang D, Zhao L, Walsh WM. Very short-term irradiance forecasting at unobserved locations using spatio-temporal kriging. Sol Energy 2015;122:1266–78. http://dx.doi.org/10.1016/j.solener.2015.10.023, URL: https://www.sciencedirect.com/science/article/pii/S0038092X15005745.

[29] Lonij VPA, Brooks AE, Cronin AD, Leuthold M, Koch K. Intra-hour forecasts of solar power production using measurements from a network of irradiance sensors. Sol Energy 2013;97:58–66. http://dx.doi.org/10.1016/j.solener.2013.08.002, URL: https://www.sciencedirect.com/science/article/pii/S0038092X13003125.

[30] Hinkelman LM. Differences between along-wind and cross-wind solar irradiance variability on small spatial scales. Sol Energy 2013;88:192–203. http://dx.doi.org/10.1016/j.solener.2012.11.011, URL: https://www.sciencedirect.com/science/article/pii/S0038092X12004021.

[31] Arias-Castro E, Kleissl J, Lave M. A Poisson model for anisotropic solar ramp rate correlations. Sol Energy 2014;101:192–202. http://dx.doi.org/10.1016/j.solener.2013.12.028, URL: https://www.sciencedirect.com/science/article/pii/S0038092X13005549.

[32] Munkhammar J, Widén J, Hinkelman LM. A copula method for simulating correlated instantaneous solar irradiance in spatial networks. Sol Energy 2017;143:10–21. http://dx.doi.org/10.1016/j.solener.2016.12.022, URL: http://www.sciencedirect.com/science/article/pii/S0038092X16306168.

[33] Shepero M, Munkhammar J, Widén J. A generative hidden Markov model of the clear-sky index. J Renew Sustain Energy 2019;11(4):043703. http://dx.doi.org/10.1063/1.5110785.

[34] Yang D. SolarData: An R package for easy access of publicly available solar datasets. Sol Energy 2018;171:A3–12. http://dx.doi.org/10.1016/j.solener.2018.06.107, URL: https://www.sciencedirect.com/science/article/pii/S0038092X18306583.

[35] Yang D. SolarData package update v1.1: R functions for easy access of Baseline Surface Radiation Network (BSRN). Sol Energy 2019;188:970–5. http://dx.doi.org/10.1016/j.solener.2019.05.068, URL: https://www.sciencedirect.com/science/article/pii/S0038092X19305493.

[36] Makarov YV, Etingov PV, Ma J, Huang Z, Subbarao K. Incorporating uncertainty of wind power generation forecast into power system operation, dispatch, and unit commitment procedures. IEEE Trans Sustain Energy 2011;2(4):433–42. http://dx.doi.org/10.1109/TSTE.2011.2159254.

[37] Lefèvre M, Oumbe A, Blanc P, Espinar B, Gschwind B, Qu Z, Wald L, Schroedter-Homscheidt M, Hoyer-Klick C, Arola A, Benedetti A, Kaiser J, Morcrette J-J. McClear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions. Atmos Meas Tech 2013;6(9):2403–18. http://dx.doi.org/10.5194/amt-6-2403-2013.

[38] Alessandrini S, Delle Monache L, Sperati S, Cervone G. An analog ensemble for short-term probabilistic solar power forecast. Appl Energy 2015;157:95–110. http://dx.doi.org/10.1016/j.apenergy.2015.08.011, URL: https://www.sciencedirect.com/science/article/pii/S0306261915009368.

[39] Wu E, Zapata MZ, Delle Monache L, Kleissl J. Observation-based analog ensemble solar forecast in coastal california. In: 2019 IEEE 46th Photovoltaic Specialists Conference (PVSC). 2019, p. 2440–4. http://dx.doi.org/10.1109/PVSC40753.2019.8980546.

[40] Yang D. Ultra-fast preselection in lasso-type spatio-temporal solar forecasting problems. Sol Energy 2018;176:788–96. http://dx.doi.org/10.1016/j.solener.2018.08.041, URL: https://www.sciencedirect.com/science/article/pii/S0038092X18308120.

[41] Yang D, Ye Z, Lim LHI, Dong Z. Very short term irradiance forecasting using the lasso. Sol Energy 2015;114:314–26. http://dx.doi.org/10.1016/j.solener.2015.01.016, URL: https://www.sciencedirect.com/science/article/pii/S0038092X15000304.

[42] Tibshirani RJ. The lasso problem and uniqueness. Electron J Stat 2013;7:1456–90. http://dx.doi.org/10.1214/13-EJS815.

[43] Tibshirani RJ, Taylor J. Degrees of freedom in lasso problems. Ann Statist 2012;40(2):1198–232. http://dx.doi.org/10.1214/12-AOS1003.

[44] Jia J, Yu B. On model selection consistency of the elastic net when $p \gg n$. Statist Sinica 2010;20(2):595–611.

[45] Yang D, Dong Z, Reindl T, Jirutitijaroen P, Walsh WM. Solar irradiance forecasting using spatio-temporal empirical kriging and vector autoregressive models with parameter shrinkage. Sol Energy 2014;103:550–62. http://dx.doi.org/10.1016/j.solener.2014.01.024, URL: https://www.sciencedirect.com/science/article/pii/S0038092X14000425.

[46] Kyung M, Gill J, Ghosh M, Casella G. Penalized regression, standard errors, and Bayesian lassos. Bayesian Anal 2010;5(2):369–411. http://dx.doi.org/10.1214/10-BA607.

[47] Tibshirani R. The lasso: some novel algorithms and applications. Purdue: Dept. of Statistics; 2011, URL: http://statweb.stanford.edu/~tibs/ftp/lassotalk.pdf.

[48] Mueen A, Zhu Y, Yeh M, Kamgar K, Viswanathan K, Gupta C, Keogh E. The fastest similarity search algorithm for time series subsequences under Euclidean distance. 2017, URL: http://www.cs.unm.edu/~mueen/FastestSimilaritySearch.html.

[49] Yang D, Alessandrini S. An ultra-fast way of searching weather analogs for renewable energy forecasting. Sol Energy 2019;185:255–61. http://dx.doi.org/10.1016/j.solener.2019.03.068, URL: https://www.sciencedirect.com/science/article/pii/S0038092X19302944.

[50] Yang D. Ultra-fast analog ensemble using kd-tree. J Renew Sustain Energy 2019;11(5):053703. http://dx.doi.org/10.1063/1.5124711.

[51] Koenker R. Quantile regression. Econometric society monographs, Cambridge University Press; 2005.

[52] Murphy AH. Climatology, persistence, and their linear combination as standards of reference in skill scores. Weather Forecast 1992;7(4):692–8. http://dx.doi.org/10.1175/1520-0434(1992)007<0692:CPATLC>2.0.CO;2.

[53] Le Gal La Salle J, David M, Lauret P. A new climatology reference model to benchmark probabilistic solar forecasts. Sol Energy 2021;223:398–414. http://dx.doi.org/10.1016/j.solener.2021.05.037, URL: https://www.sciencedirect.com/science/article/pii/S0038092X21004072.

[54] Yang D. A universal benchmarking method for probabilistic solar irradiance forecasting. Sol Energy 2019;184:410–6. http://dx.doi.org/10.1016/j.solener.2019.04.018, URL: https://www.sciencedirect.com/science/article/pii/S0038092X19303457.

[55] Yang D, van der Meer D. Post-processing in solar forecasting: Ten overarching thinking tools. Renew Sustain Energy Rev 2021;140:110735. http://dx.doi.org/10.1016/j.rser.2021.110735, URL: https://www.sciencedirect.com/science/article/pii/S1364032121000307.

[56] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Amer Statist Assoc 2007;102(477):359–78. http://dx.doi.org/10.1198/016214506000001437.

[57] Lauret P, David M, Pinson P. Verification of solar irradiance probabilistic forecasts. Sol Energy 2019;194:254–71. http://dx.doi.org/10.1016/j.solener.2019.10.041, URL: https://www.sciencedirect.com/science/article/pii/S0038092X19310382.

[58] Wang Y, Zhang N, Tan Y, Hong T, Kirschen DS, Kang C. Combining probabilistic load forecasts. IEEE Trans Smart Grid 2019;10(4):3664–74. http://dx.doi.org/10.1109/TSG.2018.2833869.

[59] Yang D, van der Meer D, Munkhammar J. Probabilistic solar forecasting benchmarks on a standardized dataset at Folsom, California. Sol Energy 2020;206:628–39. http://dx.doi.org/10.1016/j.solener.2020.05.020, URL: http://www.sciencedirect.com/science/article/pii/S0038092X20305090.

[60] Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. J R Stat Soc Ser B Stat Methodol 2007;69(2):243–68. http://dx.doi.org/10.1111/j.1467-9868.2007.00587.x.

[61] Yang D. Making reference solar forecasts with climatology, persistence, and their optimal convex combination. Sol Energy 2019;193:981–5. http://dx.doi.org/10.1016/j.solener.2019.10.006, URL: https://www.sciencedirect.com/science/article/pii/S0038092X19309880.

[62] Chen X, Du Y, Lim E, Wen H, Jiang L. Sensor network based PV power nowcasting with spatio-temporal preselection for grid-friendly control. Appl Energy 2019;255:113760. http://dx.doi.org/10.1016/j.apenergy.2019.113760, URL: https://www.sciencedirect.com/science/article/pii/S0306261919314473.

[63] Amaro e Silva R, Brito MC. Impact of network layout and time resolution on spatio-temporal solar forecasting. Sol Energy 2018;163:329–37. http://dx.doi.org/10.1016/j.solener.2018.01.095, URL: https://www.sciencedirect.com/science/article/pii/S0038092X18301166.

[64] Yang D. On adding and removing sensors in a solar irradiance monitoring network for areal forecasting and PV system performance evaluation. Sol Energy 2017;155:1417–30. http://dx.doi.org/10.1016/j.solener.2017.07.061, URL: https://www.sciencedirect.com/science/article/pii/S0038092X17306461.

[65] Hay JE, Suckling PW. An assessment of the networks for measuring and modelling solar radiation in British Columbia and adjacent areas of Western Canada. Can Geogr / Géogr Can 1979;23(3):222–38. http://dx.doi.org/10.1111/j.1541-0064.1979.tb00659.x, URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0064.1979.tb00659.x.

[66] Elsinga B, van Sark WG. Short-term peer-to-peer solar forecasting in a network of photovoltaic systems. Appl Energy 2017;206:1464–83. http://dx.doi.org/10.1016/j.apenergy.2017.09.115, URL: https://www.sciencedirect.com/science/article/pii/S0306261917314010.

[67] Lonij VP, Jayadevan VT, Brooks AE, Rodriguez JJ, Koch K, Leuthold M, Cronin AD. Forecasts of PV power output using power measurements of 80 residential PV installs. In: 2012 38th ieee photovoltaic specialists conference. 2012, p. 003300–5. http://dx.doi.org/10.1109/PVSC.2012.6318280.

[68] Engerer NA, Mills FP. KPV: A clear-sky index for photovoltaics. Sol Energy 2014;105:679–93. http://dx.doi.org/10.1016/j.solener.2014.04.019, URL: https://www.sciencedirect.com/science/article/pii/S0038092X14002151.